Identifying the impact of inframe insertions and deletions on protein function in cancer

Article (Accepted Version)

Identifying the impact of inframe insertions and deletions on protein function in cancer

Hanadi M Baeissa[1] and Frances M G Pearl[1*]

[1]Bioinformatics Group, School of Life Sciences, University of Sussex, Falmer, Brighton, BN1 9QG, United Kingdom.

*Corresponding author. F.pearl@sussex.ac.uk

Abstract

Inframe insertion and deletion mutations (indels) are commonly observed in cancer samples accounting for over 1% of all reported mutations. Few somatic inframe indels have been clinically documented as pathogenic and at present there are few tools to predict which indels drive cancer development. However, indels are a common feature of hereditary disease and several tools have been developed to predict the impact of inframe indels on protein function. In this study, we test whether six of the popular prediction tools can be adapted to test for cancer driver mutations and then develop a new algorithm (IndelRF) that discriminates between recurrent indels in known cancer genes and indels not associated with disease. IndelRF was developed to try and identify somatic, driver, and inframe indel mutations. Using a random forest classifier with 11 features, IndelRF achieved accuracies of 0.995 and 0.968 for insertion and deletion mutations, respectively. Finally, we use IndelRF to classify the inframe indel cancer mutations in the MOKCa database.

Keywords

Cancer, mutations, inframe indels

Most cancers are formed as a result of genetic mutations in DNA sequences in critical

genes that confer a selective advantage to tumour cells (Futreal et al. (2004)). These coding mutations can be caused by error in DNA replication and repair, and environmental factors that alter the genetic structure of somatic cells. Understanding the impact of these mutations is vital for providing a platform to understand cancer initiation, progression and therapeutic strategies (Hindorff et al. (2009), Ferrer-Costa et al. (2004)).

Commonly observed somatic variations in cancer include single nucleotide variants (SNV) and small insertions and deletions (Indels). Indels are the second most common type of mutations after SNVs with over two times as many deletions as insertions occurring in most cancers (Stenson et al. (2009)). Indels can affect protein function and contribute to cancer development (Akagi et al. (2010)).

Two types of indels are found in protein coding regions; frameshift and inframe mutations. Indels that cause frameshifts have a length not divisible by 3, they change the reading frame of the DNA and generally result in a change to the amino acid sequence, followed by a premature stop codon and a truncated transcript. Indels that have a length divisible by 3 are called in-frame indels and cause insertions and deletions of small runs of amino acids (Mullaney et al. (2010)).

Cancer mutations, including indels are considered driver mutations if they give the cells a selective growth advantage and contribute to the initiation or progression of the disease. Passenger mutations do not contribute to the disease progression *per se*, but occur due to the inherent genetic instability of the tumour (Greenman et al. (2007)). Driver mutations that contribute in tumorigenesis are normally found in genes described as oncogenes or tumour suppressor genes depending on their role in cancer development (Futreal et al. (2004)).

Although the majority of computational tools developed for assessing genetic mutations have focused on missense mutations, more recently, there have been several efforts to predict the impact of in-frame indel mutations on protein function or structure using a variety of strategies. Commonly used algorithms that predict the pathogenicity of impact of indels include; PROVEAN (Choi and Chan (2015)) SIFT (Hu and Ng (2013)), VEST-Indel (Douville et al. (2016)), CADD (Kircher et al. (2014)), DDIG-In (Zhao et al. (2013)), PaPI (Limongelli et al. (2015)) and PinPor (Zhang et al. (2014)).

Most of these methods classify each mutation according to two state categories; neutral or pathogenic using a variety of machine learning techniques including a J48 Decision Tree (SIFT-indel), Random Forest and Logistic Regression (PaPi) and Bayesian networks (PinPor), with reported AUC ROC accuracies varying from 0.75 to 0.9 on a variety of datasets (see Table 1).

The pathogenic mutations are generally derived from The Human Gene Mutation Database (HGMD) (Stenson et al. (2009)) a catalogue of gene lesions responsible for human inherited disease (e.g. SIFT-indel, VEST-indel, DDID-In, PaPI, PinPor) or from UniProt (PROVEAN). Neutral mutations are generally derived from the 1,000 Genomes Project (Genomes Project et al. (2010)), the Exome Sequencing Project (ESP) (Tennessen et al. (2012)) or by identifying tolerated mutations from the sequence alignment of human sequences with other mammalian species (e.g. SIFT-indel). CADD uses a slightly alternative approach that discriminates fixed or nearly fixed derived alleles in human from a set of simulated mutations. This method was developed to predict deleterious mutations rather than the functional effect on protein or variant pathogenicity using a support vector machine classifier (Kircher et al. (2014)).

In this study, rather than studying genetic mutations from model organisms and inherited disease genes, we wanted to develop a method for determining driver indel mutations specifically for somatic mutations in cancer. However, few insertion and deletion mutations have been clinically documented as pathogenic in cancer. For instance in the ClinVar database (Landrum et al. (2014)), only 20 inframe insertions and 108 inframe deletions are described as pathogenic and there even fewer reported somatic driver mutations (8 and 26, respectively).

Recurrence is often used to imply clinical driver status to cancer mutations (Landrum et al. (2014)). So to identify set of somatic indel mutations that were likely to contribute to the development of cancer we decided to use recurrence. We identified a set of recurrent somatic indels found in exome sequencing of documented cancer genes. We investigated the ability of current prediction algorithms to distinguish between these recurrent mutations and neutral indel mutations found to have little or no effect on protein function. We then defined an 'optimal' training set of cancer mutations that could be used in algorithms that predict whether an indel is contributing to the development of cancer.

An automated classifier was developed to distinguish between deleterious and neutral mutations using 11 features to describe each mutation. We selected a random forest classifier that achieved the best result to classify pathogenic and neutral mutations for insertions and deletions respectively. We validated our approach by testing our algorithm using indels clinically identified as disease causing deposited in the ClinVar database. Finally, we ran our algorithm (IndelRF) classifier to classify the predicted in-frame indels in the MOKCa database into pathogenic or neutral mutations.

Methods

Data

To identify recurrent mutations, in-frame insertions and deletions (indels) were extracted from the COSMIC database v82 using annotations form the Ensembl human genome build hg38 (Bamford et al. (2004)). Mutations were also extracted for the hg37 build of the Ensembl database for use with the PaPI, DDIG-in and PinPor algorithms.

Clinically determined cancer mutations were downloaded from the ClinVar database (Landrum et al. (2014)) with indels that were labelled as pathogenic or probably pathogenic considered pathogenic.

For the neutral set of mutation we identified a set of indels that derived from the 1000 Genomes Project and the Exome Sequencing Project (ESP) that are commonly observed in the human population (Hu and Ng (2013)). To make sure that our trained datasets were balanced, no more than 10% of the mutations within a class were taken from a single protein or a domain type.

Identification of hotspot indel mutations

To identify indels that were likely to be pathogenic, we identified hotspot mutations. For each protein in the human exome, we computed the total number of mutations it contained and the frequency of mutation at each position. A binomial test was used to identify which positions had a significant number of mutations (See supplementary methods). Insertion and deletion were tested independently and only positions where mutations occurred at least twice were analysed.

Comparison of prediction algorithms

We assessed six different algorithms that have been developed to predict the impact of in-frame indel mutations on the protein function and structure. These algorithms were: CADD (Kircher et al. (2014)), DDIG-In (Zhao et al. (2013)), PaPI (Limongelli et al. (2015)), PinPor (Zhang et al. (2014)), SIFT-indel (Hu and Ng (2013)) and VEST (Douville et al. (2016)).

Feature selection

We derived features from four existing prediction algorithms: VEST, PinPor, CADD and Pseudo Amino Acid Variant Predictor (PaPI). In total, we calculated 11 features for each mutation (See supplementary Table S1). These features describe the evolutionary conservation of the sequence where the insertion or deletion occurs, in a variety of ways, or the pathogenicity of the mutation.

Feature Importance

Mean decrease accuracy was measured to identify the variable importance using the random forest package (Archer and Kimes (2008)). The values of each of the variables in turn are randomly permuted for the out-of-bag observations, and then the modified data are passed down the tree to get new predictions. The importance of the variable is the difference in misclassification rate for the modified and original data, divided by the standard error (See supplementary methods).

Machine learning

All datasets were balanced to remove protein and domain biases in the data set. No more than 15% of mutations were allowed from a single protein or a domain family.

A random forest classifier was trained to classify pathogenic and neutral indel mutations using R version 3.2.3. Binary classifications were calculated for in-frame insertion and in-frame deletion, independently. It was run with ten fold cross validation and the parameters optimised for each model.

We also trained a support vector machine classifier (SVM) using 10-fold cross validation to optimise the hyperparameter C, used to trade off between variable minimization and margin maximization, and choose the kernel type that best fit our data.

The classifier with the best accuracy at discriminating between pathogenic and neutral mutations for both insertion and deletions was a random forest machine classifier that we have named IndelRF.

Validation of algorithms

We validated the performance of IndelRF and compared it to exisiting algorithms using test sets from ClinVar database (Landrum et al. (2016)). Predictions were generated using standard settings and the public web servers. Sensitivity (TP/TP+FN), specificity (TN/TN+FP) and accuracy (TP+TN/TP+TN+FP+FN) were measured to compare the performance of methods. We also calculated area under the curve (AUC) of receiver Operating Characteristic (ROC) curve for insertions and deletions separately.

Prediction of functional consequences of indel mutations in the MOKCa database

5437 in-frame indel mutations were downloaded from MOKCa database v21 (Richardson et al. (2009)). 1167 of them were insertions and 4270 deletion mutations.

IndelRF was used to predict whether the mutations were pathogenic and likely to be cancer drivers. We also identified the pathogenic mutations found in oncogenes and tumour suppressors as described by the Cancer Gene Census (Futreal et al. (2004)).

Results and Discussion

Identification of recurrent indels

4435 in-frame insertion mutations and 14456 in-frame deletion mutations were reported in the COSMIC database. This led to 909 recurrently mutated positions having inframe insertions and 2587 inframe deletions. As more than one indel could be reported at each amino acid position in total, there were 1856 inframe insertions and 2766 inframe mutations that we used to compare the performances of the six published algorithms.

Comparison of Prediction Algorithms

Ease of use

The number of results successfully calculated by the prediction algorithms for each of the insertion and deletion mutations, are shown in supplementary Tables S2 and S3 and Supplementary figure 1. Clearly, the algorithms did not work on all COSMIC annotations of the mutations. Often the reason was incomplete nomenclature. For instance, missing bases in the input sequences for deletions caused some algorithms to falter. The entries CTNNB1, c.14_241del228, FOXP1 c.1553_1564del12 did not give results, as the sequence of the deleted DNA was absent from the entry.

There may have also been discrepancies in genomic location of the mutation that was required for the programs due to differences in versions of the genome build used to define the mutation and that the prediction algorithm used.

Are recurrent mutations pathogenic?

In total pathogenicity values could be calculated for 898 inframe insertions and 962 inframe deletions predictions for all 6 programs available (See supplementary Figure S1). The algorithms predicted between 27%-62% insertion mutations and between 33%-73% deletion mutations as pathogenic. In total 74 inframe insertions and 109 inframe deletions mutations were predicted as pathogenic by all 6 algorithms (Figure 1). DDIG-in predicted the least number of the indels to be pathogenic whereas PaPI identified the most number of indels to be pathogenic.

Definition of optimal somatic cancer pathogenic indel datasets

To compare the variation between the algorithms, we selected 98 recurrent insertion mutations and 155 recurrent deletion mutations that had been predicted to be pathogenic by at least four of the 6 programs, as our putative pathogenic driver indel datasets. This reduction in the number of mutations was to remove protein and domain biases in the data set so that no more than 15% of mutations within a dataset were allowed from a single protein or a domain family.

When using the algorithms to distinguish between our somatic driver pathogenic indels and a neutral set of mutations, most of the algorithms performed well with accuracy scores ranging from 0.753 to 0.988, and similarly to their published performances on indels linked to hereditary disease. (see Table 1). The DDIG-in algorithm performed the best on these examples, discriminating well between the recurrent somatic cancer mutations and the neutral mutations for both in-frame indels. (Hu and Ng (2013)). The only exception was PinPor that had accuracy scores of 0.534 for insertions and 0.553 for deletions. PinPor differs to the other prediction

algorithms as it predicts the pathogenicity of indels by assessing the impact of mutations on post-transcriptional regulation rather than impact on the protein structure.

Development of a cancer specific indel classifier

Evaluation of our datasets by existing algorithms suggest that the recurrent somatic cancer mutations are pathogenic and therefore may be drivers in cancer. We then used these cancer specific datasets to train machine algorithms to enable us to detect other driver indel mutations.

Two different models, random forest and support vector machine classifiers, were trialed and compared. Binary classifications were calculated for pathogenic/neutral classes in in-frame insertion and deletion, independently.

We used a random forest classifier using 10-fold cross-validation to optimise classifier hyperparameters and assess performance for each class.

The random forest classifier has two parameters, depth and number of trees that affect on the accuracy of a classifier (Bosch et al. (2007)). The results show how the changing of both the number of trees and the depth of these trees affect the accuracy (See supplementary Tables S6 & S7) however the classification accuracy is generally high. The highest accuracy is 0.995 when the depth is 100 and the number of trees is 100 in insertion and 0.968 with a depth of 10 and 1000 tree for deletion mutations.

We also trialled a support vector machine classifier however all our random forest classifier performed better than our SVM classifier for insertion and deletion mutations. We found the highest accuracy of 0.983 and 0.962 with a radial basis function (RBF) kernel for insertion and deletion, respectively. The RBF kernel is the simplest kernel that can be used and generalizes good results (Suykens and

Vandewalle (1999, Keerthi and Lin (2003)) SVM classifier yielded the best result using RBF kernel.

The results for the SVM hyperparameter optimisation show that different values of hyperparameters in insertion and deletion mutations do not significantly change accuracy scores except when the polynomial kernel is used which caused the classifier to have a lower accuracy of 0.658 and 0.654, respectively (See supplementary Tables S8 & S9).

However, the classifier with the highest accuracy at discriminating pathogenic and neutral classes in insertions and deletions was a random forest classifier.

Feature importance

Having successfully designed an algorithm that could reliably distinguish between recurrent somatic cancer mutations and neutral insertion/deletion mutations we decided to identify the important features. Mean decrease accuracy is one of the popular feature selection methods that directly measure the effect of each feature on the accuracy of random forest. It permutes the values of one feature while others are left unchanged and measure how much the permutation reduces the accuracy (Cutler et al. (2007)).

Figure 2 shows that VEST p-value, priPhCons, Phylop and Gerp++ were the four best performing features for insertion and deletion. VEST p-value score, from VEST prediction algorithm, is the probability that benign mutation is misclassified as pathogenic. Primate PhastCons conservation score (priPhCons) was one of the top five features from CADD. Phylop and Gerp++ scores, from PaPI algorithm, are two of the evolutionary conservation score that apply different and complementary methods to weight nucleotide conservation among different species (Garber et al.

(2009)).

Moreover, the distance of indel mutation to the exon's 3' end was one of the most important features for insertions. Similarity, when comparing pathogenic versus neutral mutation for deletions, one of the top five features was the distance of indel to exon's 5' end.

Evaluation test set

We applied our algorithms to the pathogenic insertions/deletions identified in the ClinVar databases as an independent evaluation set. For somatic insertion indels, 18 pathogenic mutations and seven somatic-pathogenic mutations were evaluated using (IndelRF) with accuracies of 0.833 and 1.000, respectively. IndelRF was alos evaluated on cancer germline mutations; 72 deletion insertion 19 deletions and gave accuracies of 0.972 and 1.000, respectively. IndelRF outperformed the existing algorithms in these datasets (see Table 2).

Identifying pathogenic in-frame indel mutations in MOKCa

We applied IndelRF to the in-frame indels identified in the MOKCa database. 844 unique insertions and 1790 deletion mutations were identified. Of these (46%) 392 insertions were predicted to be pathogenic in 251 genes, and 848 (47%) deletions across 611 genes.

Analysis of pathogenic mutations

Based on the cancer gene classification in the Cancer Gene Census (Futreal et al. (2004)) we identified a set of 98 deletions in 37 oncogenes (OG) and 134 deletions across 31 tumour suppressors (TS) that were predicted to be pathogenic deletions

(see supplementary Tables S10 & S11). This suggests that indels can be both activating in oncogenes, as well as causing gene disruption in tumour suppressors.

We also detected 80 putative activating insertions across 26 oncogenes and 69 inactivating insertions across 18 tumour suppressors (See supplementary Tables S12 & S13).

Below are some of the indels predicted to be pathogenic, confirmed by reports in the literature

EGFR p.L747_E749delLRE

Epidermal growth factor receptor (EGFR) is an oncogene that regulates cell proliferation. Mutations in EGFR activate the EGFR signaling pathway and promote EGFR-mediated pro-survival and anti-apoptotic signals through down-stream targets such as RAS, RAF and MEK (Zhang et al. (2010)). The most abundant EGFR mutations are deletions in the kinase domain in exon 19 (residues 747 - 752) and constitute about 45% of all EGFR mutations (Zhang et al. (2010)). These mutations are thought to produce a conformational predisposition for the kinase to prefer its activate conformation, and hence become constitutively active.

JAK2 p.E543_D544del

Similarly Janus kinase 2 (*Jak2*) is oncogene that promotes the growth and division of cells. Jak2 mutations define a distinct myeloproliferative syndrome that affects patients with a diagnosis of polycythemia vera (PV) (Scott et al. (2007)). A small faction of polycythemia vera (PV) patients (<5%) carry usually deletions mutations in

JAK2 at exon 12 (Cazzola and Kralovics (2014, Tefferi and Pardanani (2015)) at residues E543 (Scott et al. (2007)).

KRAS p.G10_A11insG

KRAS is one of the RAS superfamily that act as oncogenes. It helps regulate cell growth. When mutated cell signaling is disrupted leading to uncontrolled cell proliferation and the development of cancer. KRAS insertion mutations have been observed between codons 10 and 11 (KRAS p.G10_A11insG) in one pateint with colorectal cancer (Tong et al. (2014)) and also in one myeloid leukaemia patient (Bollag et al. (1996)).

ARIA1A p.Q1334delQ

AT-rich interactive domain 1A (ARID1A) is a tumour suppressor that has been recognised in several types of human cancers. About 5% of ARID1A somatic mutations are in-frame indels (Guan et al. (2012)). Deletion mutations at position Q1334del were found two tumours; gastric carcinoma (Jones et al. (2012)) and prostate carcinoma (Wang et al. (2011)).

Conclusions

In this study, we sought to develop machine-learning models to identify pathogenic in-frame indels. We compared the ability of six prediction tools to discriminate between these pathogenic mutations and a set of neutral mutations, which they all did with ease.

We then developed our own classifiers that could discriminate pathogenic mutations with an accuracy of 0.995 and 0.968 for insertions and deletions, respectively. The

most four important features of our classifiers were the VEST p-value, priPhCons, Phylop and Gerp++ of in-frame insertion and deletion mutations.

Finally, we have used our algorithms to predict the functional consequence of 844 insertion mutations and 1790 deletion mutations documented in the MOKCa database.

Acknowledgements

This work was supported by King Abdulaziz University (grant number KAU1369 (to H.M.B.)).

Author disclosure statement

No competing financial interests exist.

References

AKAGI, K., STEPHENS, R. M., LI, J., et al. 2010. MouseIndelDB: a database integrating genomic indel polymorphisms that distinguish mouse strains. *Nucleic Acids Res,* 38**,** D600-606.

ARCHER, K. J. & KIMES, R. V. 2008. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis,* 52**,** 2249-2260.

BAMFORD, S., DAWSON, E., FORBES, S., et al. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer,* 91**,** 355-358.

BOLLAG, G., ADLER, F., ELMASRY, N., et al. 1996. Biochemical characterization of a novel KRAS insertion mutation from a human leukemia. *J Biol Chem,* 271**,** 32491-32494.

BOSCH, A., ZISSERMAN, A. & MUNOZ, X. 2007. Image classification using random forests and ferns. *IEEE 11th International Conference on Computer Vision,* 23**,** 1-8.

CAZZOLA, M. & KRALOVICS, R. 2014. From Janus kinase 2 to calreticulin: the clinically relevant genomic landscape of myeloproliferative neoplasms. *Blood,* 123**,** 3714-3719.

CHOI, Y. & CHAN, A. P. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics,* 31**,** 2745-2747.

CUTLER, D. R., EDWARDS, T. C., BEARD, K. H., et al. 2007. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology,* 88**,** 2783-2792.

DOUVILLE, C., MASICA, D. L., STENSON, P. D., et al. 2016. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum Mutat,* 37**,** 28-35.

FERRER-COSTA, C., OROZCO, M. & DE LA CRUZ, X. 2004. Sequence-based prediction of pathological mutations. *Proteins,* 57**,** 811-819.

FUTREAL, P. A., COIN, L., MARSHALL, M., et al. 2004. A census of human cancer genes. *Nat Rev Cancer,* 4**,** 177-183.

GARBER, M., GUTTMAN, M., CLAMP, M., et al. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics,* 25**,** i54-62.

GENOMES PROJECT, C., ABECASIS, G. R., ALTSHULER, D., et al. 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-1073.

GREENMAN, C., STEPHENS, P., SMITH, R., et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153-158.

GUAN, B., GAO, M., WU, C. H., et al. 2012. Functional analysis of in-frame indel ARID1A mutations reveals new regulatory mechanisms of its tumor suppressor functions. *Neoplasia,* 14**,** 986-993.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-9367.

HU, J. & NG, P. C. 2013. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One,* 8**,** e77940.

JONES, S., LI, M., PARSONS, D. W., et al. 2012. Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. *Hum Mutat,* 33**,** 100-103.

KEERTHI, S. S. & LIN, C.-J. 2003. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. *Neural Computation,* 15**,** 1667-1689.

KIRCHER, M., WITTEN, D. M., JAIN, P., et al. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet,* 46**,** 310-315.

LANDRUM, M. J., LEE, J. M., BENSON, M., et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res,* 44**,** D862-868.

LANDRUM, M. J., LEE, J. M., RILEY, G. R., et al. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res,* 42**,** D980-985.

LIMONGELLI, I., MARINI, S. & BELLAZZI, R. 2015. PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinformatics,* 16**,** 123.

MULLANEY, J. M., MILLS, R. E., PITTARD, W. S., et al. 2010. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet,* 19**,** R131-136.

RICHARDSON, C. J., GAO, Q., MITSOPOULOUS, C., et al. 2009. MoKCa database--mutations of kinases in cancer. *Nucleic Acids Res,* 37**,** D824-831.

SCOTT, L. M., TONG, W., LEVINE, R. L., et al. 2007. JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. *N Engl J Med,* 356**,** 459-468.

STENSON, P. D., MORT, M., BALL, E. V., et al. 2009. The Human Gene Mutation Database: 2008 update. *Genome Med,* 1**,** 13.

SUYKENS, J. & VANDEWALLE, J. 1999. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters,* 9**,** 293-300.

TEFFERI, A. & PARDANANI, A. 2015. Myeloproliferative Neoplasms: A Contemporary Review. *JAMA Oncol,* 1**,** 97-105.

TENNESSEN, J. A., BIGHAM, A. W., O'CONNOR, T. D., et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science,* 337**,** 64-69.

TONG, J. H., LUNG, R. W., SIN, F. M., et al. 2014. Characterization of rare transforming KRAS mutations in sporadic colorectal cancer. *Cancer Biol Ther,* 15**,** 768-776.

WANG, K., KAN, J., YUEN, S. T., et al. 2011. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet,* 43**,** 1219-1223.

ZHANG, X., LIN, H., ZHAO, H., et al. 2014. Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation. *Hum Mol Genet,* 23**,** 3024-3034.

ZHANG, Z., STIEGLER, A. L., BOGGON, T. J., et al. 2010. EGFR-mutated lung cancer: a paradigm of molecular oncology. *Oncotarget,* 1**,** 497-514.

ZHAO, H., YANG, Y., LIN, H., et al. 2013. DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol,* 14**,** R23.

Prediction methods	Previously published				Insertion				Deletion			
	*Sen.	*Spe.	*Acc.	*AUC	*Sen.	*Spe.	*Acc.	*AUC	*Sen.	*Spe.	*Acc.	*AUC
CADD	NA	NA	NA	0.88	0.853	0.653	0.753	0.845	0.883	0.715	0.799	0.895
DDIG-in	0.89	NA	0.83	0.89	1.00	0.976	0.988	0.991	1.00	0.936	0.967	0.975
PaPI	0.86	0.86	0.86	0.92	0.915	0.653	0.784	0.841	0.883	0.837	0.860	0.914
PinPor	NA	NA	0.75	0.83	0.830	0.238	0.534	0.533	0.680	0.389	0.534	0.553
SIFT-Indel	0.81	0.82	0.82	0.87	0.892	0.768	0.830	0.730	0.964	0.578	0.771	0.654
VEST-indel	0.90	0.90	0.90	0.91	0.923	0.700	0.811	0.886	0.982	0.872	0.927	0.973

Table 1. Comparing the performance of in-frame insertion and deletion with previously published results.
*Sen.: Sensitivity, Spe.: Specificity, Acc: Accuracy, AUC: Area under ROC curve, NA: not applicable.

	Insertion		Deletion	
	Pathogenic	**Somatic**	**Pathogenic**	**Somatic**
CADD	0.28	1.00	0.88	0.84
DDIG-in	0.56	1.00	0.86	0.84
PaPI	0.77	0.75	0.97	1.00
PinPor	0.72	1.00	0.71	0.79
SIFT-indel	0.83	1.00	0.82	0.84
VEST-indel	0.77	1.00	0.95	0.94
IndelRF	**0.83**	**1.00**	**0.97**	**1.00**

Table 2. Prediction accuracies compared between methods for four ClinVar test sets in indels.

Figures

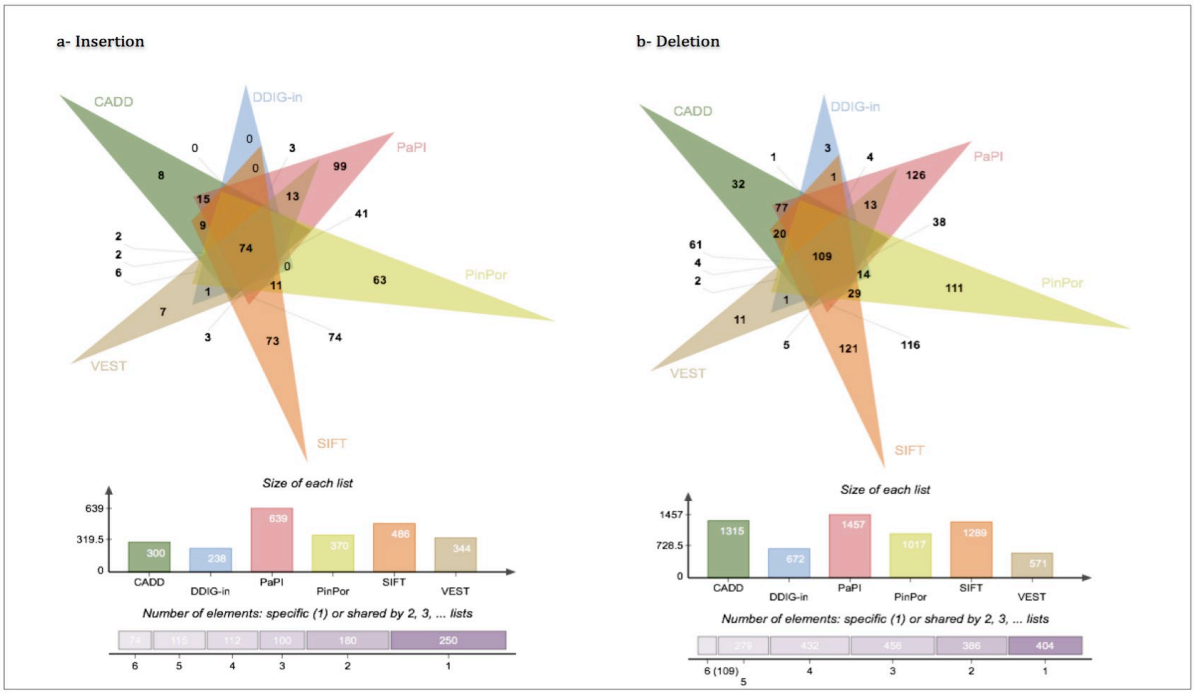

Figure 1. Common pathogenic mutations between six algorithms in inframe indels. a) in insertion. b) in Deletion

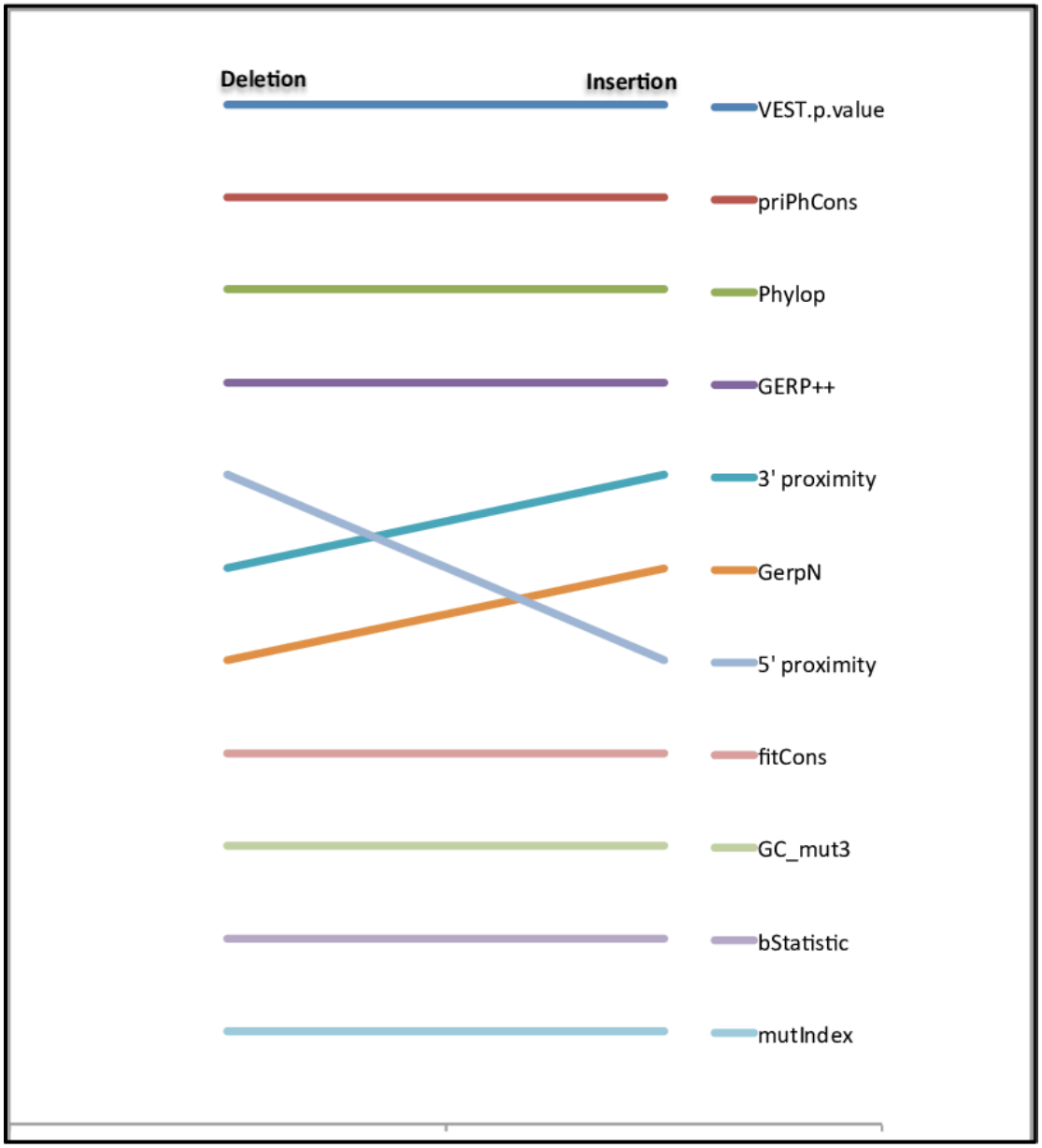

Figure 2. The importance features across insertions and deletions.
The features are ranked according to insertion mutations with the corresponding key at the side.